

영어논술평가에서의 일반 척도, 세부과업별 척도, 이분식 척도의 비교 연구

이호*·김혜영·이진화
(중앙대학교)

Lee, Ho, Kim, Heyoung, and Lee, Jin-Hwa. 2011. A Comparison of Three Rating Scales in an Argumentative L2 Essay Test: General, Task-Specific, and Binary-Choice Scale. *Korean Journal of English Language and Linguistics* 11-3, 601-626. The purpose of this study is to examine how forty student raters and two expert ones utilize each of three rating scales-general, task-specific, and binary-choice scale-in an L2 argumentative essay test. The study compared indices of inter-rater reliability across three rating scales, conducted survey analysis, and made interviews with students and experts in order to understand their perceptions on the rating scales. The results suggested that a task-specific scale yielded highest inter-rater reliability specifically in the content area. A large majority of students and all raters pointed out the task-specific scale as one which is able to measure writing competence most accurately. This paper ended up with the comment that we need to deeply explore the features and functions that a particular rating scale has in nature in order to develop the rating scale geared for the testing context for a valid and reliable writing test.

Key Words: rating scale, task-specific scale, general scale, EBB, empirically derived scale, L2 writing test

1. 서론

최근 영어 말하기·쓰기 평가와 관련된 연구가 활발해지면서 채점척도(rating scale), 혹은 평가척도표(rubric)에 관련된 관심도 높아지기에 이르렀다. 체계적으로 잘 만들어진 채점척도에 의하여 수험자의 수행능력을 측정하는 것이 평가의 신뢰도 및 타당도를 높인다는 전제하에, 현재의 평가 환경에 맞도록 기존의 척도들을 분석하

*제1저자(이호), 제2저자(김혜영), 교신저자(이진화)

여 척도를 새롭게 개발하는 사례가 늘어나고 있다. 따라서 기존 평가 척도들의 유형을 파악하고, 그 중에서 가장 적절한 유형의 척도를 파악하는 것이 중요한 평가제작 과정 중의 하나로서 인식이 되기 시작하였다(Hamp-Lyons, 1991).

평가척도의 중요성을 인식한 북미권의 학자들은 다양한 분류기준에 의하여 평가척도의 유형들을 제시하고 각 평가척도들의 신뢰성과 타당도를 논의해 오고 있다(Carr, 2000; Lane, 2008; Moskal, 2000). 척도를 분류하는데 있어서, 학자들은 평가대상 영역에 따라서 세부척도(analytic scale)와 총괄척도(holistic scale)로, 전반적 점수 결정 과정에 따라서 관련 평가 영역 점수를 합산하여 총점을 매기는 총합평가방식(composite scoring)과 수험자의 전반적인 능력에 대한 채점자의 인식에 의존하여 결정을 내리는 총괄평가(holistic scoring)로, 척도개발 방식에 따라서 직관적/연역적 척도 및 실증적/귀납적 척도 등으로 다양하게 세분화하였다(East, 2009). 이렇게 다양한 평가척도 유형들은 교사나 평가관계자가 평가환경에 맞는 적절한 평가척도를 선택하는데 참고가 되고 있다.

하지만, 특별히 EFL상황에서 영어 평가척도의 유형이나 기능을 다룬 지침서나 논문은 드문 현실이다. 다수의 논문들은 특정 평가 영역 측정의 타당성을 논의하는 과정 중의 하나로 연구 중에 사용된 척도에 대하여 단순히 소개하거나, 단일 척도에 대한 개발 과정을 설명하는 데 그칠 뿐, 척도별 상대적 비교를 통하여 각 척도의 기능과 특색을 심도 깊게 논의한 논문은 찾아보기 어렵다. 전체 수행평가의 신뢰도와 타당도에 있어서 평가 척도가 미치는 영향을 상기해 본다면, 영어 평가 척도에 대한 깊이 있는 탐색이 필요한 것은 매우 자명하다.

본 연구는 영어쓰기 평가분야에서 척도 타당성에 대한 논의의 중요성을 제기하면서, 다양한 유형의 평가척도가 연구 참여자 집단에 어떻게 작용하는지 보고자 한다. 특히 추후 학교 현장에서 평가를 담당해야 할 예비교원들을 수험자이자 채점자로 참여시켜서, 예비교원들의 평가전문성 함양의 효과도 갖고자 한다. 본 논문에서는 영어평가 상황에서 보편적으로 사용되는 직관적/연역적 방식의 일반척도와 개별 문항 측정에 적합한 실증적/귀납적 방식의 세부과업별/개별 척도를 대조하여 살펴보고, 아울러 요즘 새롭게 주목을 받

고 있는 이분식 척도까지 포함하여 척도별 비교 연구를 수행할 것이다. 따라서 본 논문은 크게 다음의 세 가지 연구 질문을 수립하였다.

1. 각 척도에 걸쳐 채점자의 평가 양상은 어떻게 나타나는가?
2. 채점자로서의 수험자는 각 척도에 대하여 어떻게 인식하는가?
3. 전문인 채점자 측면에서 각 척도에 대한 인식은 어떠한가?

첫 번째 연구 질문은 정량적 연구 방법인 급간상관계수 및 ANOVA를 통하여 신뢰도를 파악하고, 두 번째 연구 질문은 설문지 분석과 인터뷰 분석을 병행하여 실시하며, 마지막 연구 질문은 사고표출기술기법(think-aloud protocol analysis)을 통한 채점과정 탐색과 인터뷰를 통한 채점척도에 대한 채점자 인식 조사를 수행할 것이다. 전반적으로, 혼합연구방법을 채택하되, 정성적 연구방법과 정량적 연구방법을 동등한 비중으로 수행하였다.

2. 문헌 연구

평가척도의 유형과 관련하여 빈번하게 언급되는 분류는 총괄척도(holistic scale)와 세부척도(analytic scale)이다(East, 2009; Weigle, 2002). 총괄척도는 독자가 책을 읽고 평가할 때의 자연적인 반응을 평가적 상황에 구현한다는 것으로, 학습자의 전반적인 능력은 채점자의 인상적 평가에 의하여 단일점수로 표현된다. 이 총괄척도는 주로 상대평가에 의하여 학생들의 점수를 결정하는 총괄적 평가에서 자주 사용된다(Charney, 1984). 세부척도는 다양한 세부영역에 걸쳐 학습자의 능력을 종합적으로 측정하며, 높은 신뢰도와 구인타당도(construct validity)를 지니고 있다. 이러한 세부척도는 개발하기에 시간과 비용이 많이 들지만, 학습자의 정보를 자세하게 알려주기에 진단평가적 성격에 적합한 척도표라고 할 수 있다(Knoch, 2009; Weigle, 2002).

또 다른 평가척도의 분류로는 평가문항과 척도와의 관계에 따라 일반적 척도(general scale)와 세부과업별/개별척도(task-specific scale)로 나눌

수 있다(Zimmerman, 2004). 기존의 TWE(Test of Written English)와 같은 대규모의 고부담시험에서는 하나의 척도를 가지고 다양한 평가과업을 총괄점으로 평가하는 경우가 많은데(McGriff, 2007, Weigle, 2002), 이와 같이 하나의 척도를 모든 과업에 공통적으로 적용하는 경우가 일반적이라고 할 수 있다.

세부과업별 척도(task-specific scale)는 문항마다 혹은 과업마다 측정하는 평가기준들이 다른 것이 특색이다(Moskal, 2000). 세부과업별 척도는 채점자간 신뢰성이 높아서 채점의 일관성이 중요하게 취급되는 특수목적의 고부담시험(high-stakes test)에서 사용될 수 있다(McGriff, 2007). 여기서 특수목적의 시험이란 측정해야 할 단면을 두 가지 이상 갖고 있는 다차원성(multi-dimensionality)을 특색으로 하며, 언어적 능력이외에 특정영역에 대한 지식의 완비 여부를 중요하게 측정하는 시험이라 할 수 있다(Douglas, 2000). 만약에 이런 특수 목적의 수행평가의 상황에서 한 개 이상의 과업이 주어지고 각 과업마다 측정하고자 하는 내용 영역이 각각 다른 경우에는, 각 문항에 특화된 세부 과업별 척도표가 적합하다고 할 수 있다. 반면, 이러한 지식의 모든 과업에 공통적으로 적용하기 어렵고 언어학습자의 일반적인 언어적 능력을 측정하기가 어렵다는 단점이 존재한다(Chalhoub-Deville, 1995).

마지막으로 학자들 사이에 많이 제시되고 있는 분류 방식은 척도 개발 방법에 의한 직관적/연역적 척도(intuitively derived scale)와 실증적/귀납적 척도(empirically derived scale)의 분류이다(Condon, 2009; Fulcher, 1996; Knoch, 2009). 기존의 일반적 척도는 대부분 직관적/연역적 척도로서, 전문가들이 기존 평가척도들을 분석하고 전문가간 상호 협의 과정을 거치면서 본인의 직관적 경험에 의하여 척도를 만드는 경우가 많았다(Fulcher, 1996). 최근 들어서는, 수험자들의 실제 발화자료나 쓰기자료의 분석을 통해, 특정 평가환경에 적합한 세밀한 평가영역과 평가기준을 제작하는 실증적/귀납적 척도가 각광을 받기 시작하였다(Condon, 2009; Knoch, 2009; Upshur & Turner, 1999). Knoch(2009)에 따르면, 실증적/귀납적 척도는 그 상대편 척도에 비하여 채점자간 신뢰도가 더 높고, 채점자들에게 쓰기의 다양한 측면들을 더 잘 분별해준다고 보고하였다.

이런 실증적/귀납적 척도의 대표적인 것으로 귀납적 이분법적 평가

척도(EBB: Empirically derived, Binary-choice, Boundary-definition)를 들 수 있다. EBB는 다양한 평가항목들을 중요도에 따라 단계별로 서열화하고 채점자가 각 단계를 진행하면서 각 평가항목마다 이분법적인 선택을 하는 체제이다(Upshur & Turner, 1995). 이 EBB는 말하기 평가 상황에서 주로 사용되었으며, 다수의 논문들은 EBB가 진단적 평가에 적합함을 주장하고 있다(Upshur & Turner, 1999; 황성삼, 2008).

말하기 평가의 경우 이분식 평가 척도를 활용한 다수의 연구들이 발표되어 왔다. 외국의 사례를 보면, Upshur와 Turner(1995)의 논문에서 EBB를 최초로 활용한 논문을 발표하였는데, 그들은 초등학생들의 이야기 서술 능력을 EBB를 활용하여 측정하였다. 그들은 EBB의 특성으로서 평이함과 명확성을 제시하고 채점자간의 신뢰도가 매우 높았다는 결과를 발표하였다. 그들의 추후 연구(Upshur & Turner, 1999)에서도 데이터에 기반한 귀납적 방식으로서의 EBB에 대한 긍정적 측면들을 서술하였다.

한국의 EFL상황에서도 말하기 평가에서의 EBB활용을 다룬 논문들이 소수 나왔다. 중학생의 말하기 총괄평가 상황에서 EBB를 활용했던 황성삼(2008), 고등학생의 말하기 총괄능력을 측정한 Song(2009), 예비 영어교사로서의 대학생들의 평가능력을 탐색한 이호(2010)의 연구 등이 발표되었다. 이 학자들은 공통적으로 EBB척도의 높은 신뢰도와 실용성, 교육 현장과의 용이한 연계성 등을 장점으로 제안함과 동시에 EBB척도의 변별력의 문제와 척도 개발의 어려움을 보고하였다.

반면, 쓰기평가 분야에서는 EBB와 같은 이분식 척도를 활용한 논문이 매우 드문 실정에서, Kim(2010)의 연구가 주목할 만하다. 그녀는 이분법적 평가 척도인 귀납적 평가항목기반 진단 대조표(EDD: Empirically-derived Descriptor-based Diagnostic checklist)를 쓰기평가 측정에 도입하였다. 그녀는 이분식 평가 척도를 제작하는 과정에서 다양한 통계적 방식과 정성적 연구기법을 결합하여 척도표의 신뢰성과 타당성을 고양하였다. 9명의 ESL 채점자들로부터 사 고표출구술기법을 통하여 35개의 이분식 평가 항목을 도출하였고 10명의 ESL 교사들이 이 평가 척도를 활용하여 480개의 ESL학생들의 에세이를 평가하였다. 그 채점 자료를 바탕으로, 연구자는 DIMTEST나 DETECT, CCPROX/HCA와 같은 다면도 측정(dimensionality analyses) 프로그램을 활용하여 이분식 평가 척도가 크게 다섯 개의 단면도

(내용수행, 구조의 효율성, 문법지식, 어휘사용, 형식적 규칙-mechanics)를 포함하고 있음을 밝혔고, EDD의 사후 검증과정으로서 환원법적 관련인수 통합 모형(the reduced reparameterized unified model)을 통하여 각 평가항목의 진단적 기능을 검증하였다.

그녀의 연구는 몇 가지 측면에서 의미가 있다. 첫째로, 기존의 말하기평가에 치중되었던 이분법적 척도의 활용을 작문 평가 상황에서도 시도하였다는 점이다. 둘째로, 척도의 신뢰도 및 타당도를 논의하는 각 단계마다, 정성적 연구와 정량적 연구를 거의 동일한 비중으로 적절하게 수행하여 논의의 폭과 깊이를 심화시켰다는 점이다. 마지막으로, 읽기와 듣기평가의 사지선다형 문항 분석에서 주로 다양하게 사용되었던 심리측정학적 기법이 수행평가인 작문평가에서도 통용될 수 있도록 했다는 점이다.

지금까지, 다양한 척도의 유형에 대하여 정리하였다. 기존의 많은 연구들이 제안한 것은 각 척도마다 장단점이 존재한다는 것이고 따라서 특정 척도가 다른 척도에 비하여 절대적인 타당도가 높다고 단언하지 못한다는 것이다. 그러나, 그러한 전제에도 불구하고 실제 개발 단계에 있어서는 척도별 비교를 통하여 채점자와 평가개발자에게 가장 적합한 척도를 선택하는 것이 필요하다는 점을 생각할 때, 실제 교육 현장에 유용한 정보를 제공할 수 있는 척도별 비교 연구가 수행되어야 할 것이다.

3. 연구 방법

3.1 연구 참여자

연구 참여자들은 서울의 C사범대학에 재학 중인 46명의 학생들로서 영어학습자임과 동시에 추후 학교 현장에서 중·고등학교 학생들에게 평가를 수행할 예비교원들이다. 참여자의 성별은 남자가 11명, 여자가 35명으로 여자가 대다수이며 참여자들의 전공은 거의 영어교육 전공자들이다. 학기초에 실시된 영어논술시험 결과를 토대로 수험자들은 대부분 중상급이상의 영어쓰기 능력을 보유하고 있음을 확인하였다. 수험자들은 간단한 평가자 훈련과 함께, 모두 세 차례에 걸쳐 영어논술시험에 수험자로 참여하였으며, 자기평가

와 동료평가를 수행하는 등 채점자로서도 역할을 하였다. 46명 중 에세이를 늦게 제출한 6명을 제외한 40명의 3차 영어논술 시험 자료를 연구질문 1의 분석 대상으로 삼았다. 하지만 설문지에는 46명 전원 참여하였다. 추후 연구자는 무작위로 7명의 학생들을 선정하여 개인당 30분간의 인터뷰를 실시하였다. 마지막으로 본 연구에 참여한 채점자 전문가 2인은 영어교육분야에서 석사 학위 이상을 수여받은 대학원생으로서, 두 명 모두 EFL 환경에서 학생들에게 영어를 가르친 경험이 있고 대학교의 교양영어 말하기·쓰기 평가를 다수 수행하여 평가자로서 어느 정도의 전문성을 갖추고 있다고 할 수 있다. 2명 모두 본 연구를 위하여 각 척도별로 2시간 이상의 평가자 훈련 과정을 수행하였다.

3.2 연구 자료

3.2.1 시험 문항

본 연구 참여자들의 대부분이 임용고시를 준비하고 있는 학생임을 고려하여, 연구자들은 중등교사 임용고시 2차 논술형 형식에 따라 시험문항을 제작하였다. 3차례의 영어논술 시험문항들은 각기 '영어교육', '영어학과 영문학', '일반영어'의 주제로서 출제되었으며 본 연구의 분석대상이 되는 3차 시험 문항 주제는 '일반영어' 이다. 3차 논술문항은 국제결혼이 단일 국가 개념에 미치는 영향에 대하여 서술하는 것으로 서론·본론·결론의 3개 문단 이상을 구성하여 작성하도록 하였다.

3.2.2 3개의 척도표: 일반척도표, 세부과업별 척도표, 이분식 척도표

본 연구를 위하여, 연구자들은 3개의 척도표(일반척도표, 세부과업별/개별척도표, 이분식척도표)를 사용하였다. 세부과업별/개별척도표는 예전 영어작문 수업시 활용했던 척도를 학생들의 설문 의견을 반영하여 본 논술문항에 맞게 수정·개발하였으며, 평가영역은 크게 '내용'과 '구조 및 어휘사용'으로 나누었다. 세부채점요소로서 모두 아홉개의 요소를 두었으며 아홉 개 요소의 각 부분별 총합 25점 만점 체제로 개발하였다(부록2 참조). 아홉 개 요소 중 세 가지 평가항목은 내용의 영역으로, '서본결의 구조가 제대로 이루어져 있다'라는 평가항목을 구조의 영역으로, 나머지 네 개의 평가 항목은

언어사용의 영역으로 결정하였다. 본 연구에서는 '300자 내외로 작성되었다'는 평가 항목을 총합점수에는 포함시키고 내용점수합계에서는 누락시켰는데, 그 이유로는 이 평가항목이 다른 내용 평가 항목들과는 달리 에세이 내용의 질을 평가하기 보다는 기계적으로 에세이의 분량을 측정하고자 했기 때문이다.

이분식 척도표는 Kim(2010)의 박사논문에서 사용된 귀납적 평가 항목기반 진단대조표를 그대로 사용하였다(부록3 참조). 이분식 척도표는 총 35개의 문항으로 구성되어 있으며, 모든 문항은 Yes 혹은 No의 두 등급으로 구성되어 있어서, 만점은 35점 체제이다. 그녀의 사후분석에 의하면, 1번부터 8번까지의 8개 문항은 내용영역에 해당되는 지표이고, 9번부터 14번까지의 6개 문항은 구조영역에 해당되는 지표로 분류를 하였다. 나머지 15번부터 35번까지 21문항은 문법, 어휘, 형식적 규칙에 해당되는 것으로 모두 언어사용과 관계된 것이었다. 본 논문에서는 이 분석을 그대로 인용하여 이분식 척도표의 세부평가 영역을 내용, 구조, 언어사용 세 가지로 분류하였다.

일반척도표(부록1 참조)는 지난 2년동안 영어작문 수업을 위해 사용되었던 척도표를 본 평가 환경에 맞게 약간 수정하였다. 일반척도표는 모두 3개의 명시적인 세부영역(구조, 내용, 어휘사용)으로 구성되어 있으며, 다른 두 가지의 척도표들과는 달리 총괄점수 영역을 배치하였다. 총괄영역과 세부영역은 각 9점 등급 척도로 하였다. 그밖에 각 세부영역 안에 5등급제의 10개의 정밀채점영역을 배치하였다(부록1 참조).

3.2.3 설문지

각 평가척도의 특성에 관한 학습자들의 인식을 알아보기 위해 설문지 조사를 시행하였다. 설문지는 모두 18개의 문항으로 작성되어 있으며, 본 연구주제와 관련있는 6개의 설문 문항을 추출하여 분석하였다. 그 중 3개의 문항은 삼지선다형 문항으로 '제일 채점하기 편리했던 척도표', '자신의 영작실력을 향상시키는 데 있어서 가장 도움이 될 것이라 생각하는 척도표', '비교적 정확한 채점이 가능한 척도표'에 3가지의 척도표 중 해당되는 척도표를 선택하도록 하였다. 나머지 3개 문항은 개방형 문항으로서 각 척도표의 장단점을 서술

하도록 하였다.

3.3 자료수집과정

연구 참여 학생들은 3차례의 영어 논술 모의시험을 보기 전에, 두 시간 정도의 채점자 훈련 과정을 거쳤다. 연구 참여 학생들은 3가지의 척도표에서 제시된 각각의 평가 항목들에 대한 설명을 듣고, 각 척도표를 가지고 모의 채점을 진행하였다. 연구 참여 학생들 중 다수는 이미 기존 영어작문 수업에서 사용되었던 일반적 척도에 대하여 익숙하므로, 연구자는 개별척도와 이분식척도에 대하여 집중적으로 안내를 하였다. 채점자 훈련 과정이 끝난 후, 학생들은 세 차례의 영어 논술 시험을 치루고, 각 시험마다 세 가지 척도를 가지고 자기평가와 동료평가를 수행하였다. 마지막 3차 시험이 끝난 이후, 학생들은 설문지 응답에 참여하였다.

연구자는 추후 7명의 학생들을 인터뷰 참여자로 선정하여 사무실에서 일대일 면담을 수행하였다. 면담은 반구조적 형태(semi-structured interview)로 진행되었으며, 학습자의 반응에 따라서 면담 질문을 조금씩 수정하기도 하였다. 면담은 회상자극기법(simulated-recall technique)을 활용하였는데, 학습자들이 작성해서 제출했던 점수 자료와 설문지 자료를 학습자들의 구술 응답과 비교해 가며 이루어졌다.

모든 영어 논술 시험이 끝난 이후, 두 명의 전문인 평가자가 학생들에게 의하여 작성된 3차 영어논술 에세이들에 대하여 채점하였다. 채점자들에게도 학생들과 마찬가지로 3가지의 척도가 주어지고 그 척도에 대한 간단한 워크샵을 가진 이후 에세이에 대하여 평가를 하도록 하였다.

3.4 자료 분석

연구자는 혼합연구방법(mixed-method approach)을 채용하여 연구를 진행하였다. 연구질문 1의 분석은 정량적 방식, 연구질문 2에 대한 분석은 정성적 방식과 정량적 방식을 모두 사용하였으며, 연구질문 3의 분석을 위해 정성적 방식을 채택하되 추가적으로 연구질문 1에서 발견된 현상을 검증하고 검토해 보았다. Dörnyei(2007)의 지침에 따라서, 연구자는 정성적 연구와 정량적 연구가 상호 보완적 관계로 동등한 비중으로 수행되도록 하였다.

3.4.1 연구 질문 1의 자료 분석: 기술통계, 성취율과 급내상관계수

연구질문 1에서는 40명의 수험자가 작성한 에세이에 대하여 학생 평가(자기평가, 동료평가)와 전문인 평가(전문인1 평가, 전문인2 평가)가 3개의 척도(일반척도, 세부과업별/개별척도, 이분식척도)를 기반으로 하여 매긴 구조, 내용, 어휘사용, 총합 점수들을 분석대상으로 하였다. 일반척도에 한해서는 에세이의 전반적인 질을 채점자의 인상에 의하여 평가하는 총괄 점수를 추가하였다. 연구자는 기술통계를 사용하여 각 평가자 영역별/척도별 평균과 표준편차를 계산하였다. 평균은 각 평가자 그룹의 영역별/척도별 채점의 관용도(rating severity)를 비교할 수 있고, 표준편차는 각 평가자별로 채점 점수의 변이도를 파악할 수 있었다. 성취율은 등급 체제가 각기 다른 척도간의 점수들을 동등하게 비교하기 위하여 본 연구자가 계산한 것인데, 그 공식은 '(해당척도의 특정영역 평가자들 점수의 평균/해당척도의 만점) × 100'이다. 즉, 만점이 각각 다른 척도들을 비율로 변환하여 동등하게 상호 비교할 수 있도록 했다. 성취율이 높다면, 해당 영역에 대하여 학생들의 성적이 우수하거나 채점자들이 후한 평가를 했다는 의미가 된다. 마지막으로 연구자는 각 척도의 채점자간 신뢰도를 비교하기 위하여 급내상관계수(ICC: Intra-Class Correlation)를 적용하였다. 급내상관계수는 두 명이상의 채점자가 포함된 수행 평가 상황에서 적용 가능하며, 실험집단의 수가 적은 반복측정의 상황에서 이상적으로 사용할 수 있는 통계적 기법이다. 보통은 급내상관계수의 평균측도가 자주 쓰이며, 이 평균측도 값은 Cronbach의 알파값과 일치한다(Hwang, 2010).

3.4.2 연구 질문 2의 자료 분석: 설문문항분석과 인터뷰

연구 질문 2를 위하여 설문지에 참여한 46명의 학생들의 삼지선 다형 문항과 개방형 문항 분석을 병행 실시하였다. 면담 분석과 관련하여, 연구자는 7명과의 면담내용을 녹음한 후 이를 모두 문서화하였다. 연구자들은 문서화된 면담내용과 연구 질문 1에서 사용되었던 정량적 자료들을 면밀히 대조하여 살펴본 후, 주요 면담 내용 자료를 분석틀에 맞게 발췌하는 작업을 수행하였다.

3.4.3 연구 질문 3의 자료 분석: 사고표출구술기법과 인터뷰

연구 질문 3을 위해서 연구자는 사고표출구술기법을 통하여 채점자의 채점 과정을 분석하고자 하였다. 사고표출구술기법은 평가자가 채점하는 과정 중에서 본인의 마음속에 있는 표현들을 모두 구술로 표출하는 것(Green, 1998)으로 본 연구에서는 통시적(concurrent) 방식을 이용하여 채점 상황 그 당시의 채점과정을 관찰해 보고자 하였다. 사고표출기법을 통해 나온 데이터는 추후 전문 채점인과의 면담 내용과 교차 비교하여 그 데이터의 신뢰도를 확인하는 삼각구도방법론(methodological triangulation)을 시행하였다.

4. 결과 및 논의

4.1 각 척도에 걸쳐 채점자의 평가 양상은 어떻게 나타나는가?

표 1은 각 평가자가 매긴 점수의 영역별/척도별 평균과 표준편차를 제시한 것이다. 평균값에 의하면, 척도에 상관없이 가장 관용적인 채점행태를 보인 평가자그룹은 동료평가라고 할 수 있다. 반면, 가장 혹독한 채점경향을 보인 채점자는 모든 척도에 걸쳐서 전문인 2라고 할 수 있다. 전반적으로 학생평가자보다 전문인 평가자들이 다소 비관용적인 채점행태를 보이고 있었다.

표 1. 각 평가자 점수 영역별/척도별 평균 비교

척도	평가자	총괄(SD)	내용(SD)	구조(SD)	언어(SD)	총합(SD) ⁴
일반 ¹ (N=40)	자기평가	5.70(1.31)	5.73(1.34)	5.82(1.68)	5.78(1.61)	17.33(4.05)
	동료평가	6.37(1.28)	6.05(1.30)	6.33(1.58)	6.53(1.28)	18.90(3.65)
	전문인1	4.88(2.77)	5.60(2.64)	5.92(2.82)	5.60(2.80)	17.13(7.79)
	전문인2	5.08(1.14)	5.10(1.30)	5.12(1.16)	4.88(1.14)	15.10(3.34)
개별 ² (N=40)	자기평가	-	7.20(1.56)	2.43(.68)	8.33(1.40)	19.72(3.03)
	동료평가	-	7.53(1.11)	2.55(.60)	8.97(1.25)	20.80(2.39)
	전문인1	-	6.88(1.92)	2.38(.77)	8.15(2.61)	19.20(4.65)
	전문인2	-	6.43(1.48)	2.25(.63)	7.73(1.28)	18.32(2.90)
이분 ³ (N=40)	자기평가	-	4.93(1.94)	4.34(1.46)	16.22(3.89)	25.49(6.01)
	동료평가	-	5.78(1.78)	4.37(1.45)	16.59(3.29)	27.51(7.44)
	전문인1	-	5.02(2.76)	4.37(2.11)	15.39(5.45)	24.78(9.30)
	전문인2	-	4.34(2.37)	4.22(1.37)	15.37(3.45)	23.93(6.35)

Note. 일반¹척도의 총괄, 내용, 구조, 언어 영역 만점은 각각 9점, 총합 27점
 개별²척도의 만점은 각각 내용 9점, 구조 3점, 언어 11점, 총합 25점
 이분³척도의 만점은 각각 내용 8점, 구조 6점, 언어 21점, 총합 35점

각 평가자마다 제시되어 있는 표준편차 값은 채점자의 채점전략에 대해서 어느 정도 정보를 제공하고 있기에 자세히 살펴 볼 필요가 있다. 상대적으로 표준편차가 큰 평가자는 상위학습자와 하위 학습자간의 점수 간 차이를 명확하게 분류하여 채점했을 가능성이 있고, 이와는 반대로 표준편차가 적은 채점자는 학생들 간의 점수 차이를 다소 억제하여 평가했다는 점을 알 수 있다. 여기서, 전문인 1의 점수자료에서는 다른 채점자에 비하여 수험자간 점수의 차이 폭이 크다. 반면에, 전문인2는 수험자간 점수가 적게 평가를 했고 이는 극단의 점수를 피하는 안전지향적 채점 전략을 차용했을 가능성이 있다는 것이다.

표 2는 평가척도간의 채점자들의 채점경향 비교를 위하여, 각 척도별 만점대비 수험자의 성취율을 제시하였다. 세 척도의 성취율을 총합점수 측면에서 상호 비교하면, 채점자들은 개별/세부과업별 척도에서 가장 후한 점수를 준 반면 일반척도에서 다소 냉혹한 점수를 준 것으로 해석된다. 영역별로 보면, 내용 영역에서 상대적으로 가장 낮은 점수를 기록하였고 구조가 가장 높은 점수를 받았다고 할 수 있다.

표 2. 각 평가척도 영역별 성취율 비교

척도	총괄	내용	구조	언어	총합
일반 (N=40)	61.19	62.44	64.42	63.31	63.39
개별 (N=40)	-	77.89	80.08	75.41	78.04
이분 (N=40)	-	62.72	72.08	75.68	72.65
평균	61.19	67.68	72.19	71.47	71.36

표 3은 급내상관계수에 의한 채점자간 신뢰도를 제시하였다. 총합영역 측면에서 모든 채점자를 고려할 때, 채점자간 신뢰도가 가장 높았던 척도표는 세부과업별/개별척도표(.707)였다. 이 척도표는 전문인과 학생에서 모두 제일 높은 신뢰도를 기록하였고, 특히 내용영역(.705)에서 채점자간 신뢰도가 높게 나왔다. 반면에, 일반척도의 총괄영역은 세부영역 점수들을 합산하여 총점을 계산하는 총합적 방식에 비하여 신뢰도 수치가 매우 낮게 나와서, 본 시험환경에서는 다소 부적합한 평가체제로서 판단된다. 전반적으로 학습자들에 의한 동료-자기평가간 평가자 신뢰도가 전문인 간 평가자 신뢰도에 비하여 높이 나온 점이 주목할 만 하다.

표 3. 급내상관계수(ICC)에 의한 채점자간 신뢰도

척도	평가그룹	총괄	내용	구조	언어	총합
일반 (N=40)	동료-자기	.548	.648	.697	.771	.749
	전문인 간	.420	.396	.441	.411	.476
	채점자합산	.503	.476	.658	.653	.648
개별 (N=40)	동료-자기	-	.706	.591	.625	.752
	전문인 간	-	.619	.645	.553	.664
	채점자합산	-	.705	.650	.610	.707
이분 (N=40)	동료-자기	-	.653	.684	.766	.660
	전문인 간	-	.538	.565	.392	.472
	채점자합산	-	.656	.684	.660	.679

4.2 학습자 측면에서 각 척도에 대한 인식은 어떠한가?

표 4는 설문지 조사를 토대로, 학습자들이 각 평가척도에 대하여 어떤 인식을 갖고 있는지를 상호 비교하였다. 다수의 학생들(46명중 23명)은 채점하기 용이한 척도표로서 이분식 척도를 꼽았고, 반면 영어작문 학습에 도움이 되거나(46명중 28명) 가장 정확하게 채점을 할 수 있는(46명중 36명) 척도표로서 개별/세부과업별 척도를 지적하였다.

표 4. 각 척도표 사용에 관한 학습자의 응답분포

평가척도	채점하기 편리	학습에 도움	정확한 채점 가능
일반척도	10	4	3
개별척도	13	28	36
이분척도	23	14	7
합계	46	46	46

설문지 개방형 문항 분석은 학생들이 생각하는 각 척도에 대한 장단점을 제시하고 있다. 학생들은 일반식 척도의 장점으로 대체적으로 점수 선택의 폭이 넓어 정확한 수치 정보가 제공된다는 점, 총괄영역과 세부영역을 연결하여 유기적으로 점수를 줄 수 있고 융통성 있는 채점이 가능하다는 점 등을 들었고 단점으로는 특정주제에 대한 세부적인 평가의 어려움, 점수별 구분의 모호성, 채점 기준표를 일일이 대조해서 봐야하는 번거로움 등을 호소하였다.

세부과업/개별척도에 대하여 학생들은 그 장점으로 해당주제에 대한 평가의 정확성과 높은 신뢰도, 평가문항을 이해하기 쉬운 점 등을 제시하였다. 반면에 학생들은 문항에서 요구하는 부분만 평가

하기 때문에 전반적인 학생의 작문능력을 자세히 평가하기 어렵다는 점, 등급이 3등급으로만 되어 있어서 변별력이 떨어질 수 있는 점, 세부적인 측정에 매달리다 보면 전체적인 그림을 놓칠 수 있다는 점 등을 대표적인 단점으로 지적하였다. 이분식 척도의 경우, 대부분의 학생들은 가장 편리하고 효율적인 쓰기 척도라는 점을 강조하고 있고 반면에 Yes나 No의 단 두 가지 선택의 강요가 중간 정도의 실력을 측정하는데 어려움을 주고 있다고 호소하였다.

이분식 척도의 추가적인 장점에 대하여 ‘박’은 이분식 척도가 진단적 평가 목적으로서 적합할 수 있음을 지적하였다. 35개나 되는 항목들을 평가해 가면서 자신의 영어 실력을 구체적으로 검증할 수 있는 계기가 되었다는 ‘박’은 아래와 같이 덧붙여서 언급하였다.

실력향상을 하는데 가장 도움이 되는 건.. 아무래도 이거(이분식 척도)요. 아무래도.. 자세하니까.. 뭘 잘했는지 뭐가 잘못됐는지 알 수 있으니까. 문항수가 많아서 자신에 대하여 구체적으로 잘 알 수도 있고. 대신에 문항 수가 많아서 평가하기 좀 힘들어서 그렇지. 이게 스탠다드 잡기도 그렇고. (박, 인터뷰)

그러나, 이런 ‘박’의 경우에도 이분식 척도의 문제점을 평가할 항목이 많아 어렵다는 것으로 언급하였는데, 아래 인터뷰 참여자들은 이분식 척도의 등급설정의 문제점을 지적하였다.

솔직히 제가 제일 어려웠던 거는 Yes/No 인데요. 그 이유는 딱 한 문장만 가지고서는 전체를 평가할 수 없잖아요. 이런 부분도 있지만 그렇지 않은 부분도 있는데..그럴 경우에는 이것을 Yes라고 해야 할지 No라고 해야될지를 판단이 잘 안서니까. 저 같은 경우에는 Yes/No가 제일 어려웠고. (유, 인터뷰)

이분척도평가는 너무 딱 Yes/No여가지고요.. 사실은 저는 항상 저는 중간 정도 수준이라고 생각을 해서 맨 마지막 그 이분식 척도 평가는 좀 정확하게 채점이 안됐던 것 같구요. (김, 인터뷰)

위에서 언급한 이분식 척도에서 나타난 등급의 협소함과 선명히 대비하여 등급의 다양함을 일반식 척도의 큰 장점으로 제시하는 참여자도 있었다. ‘노’는 등급이 많다는 것이 세밀한 평가를 가능하게 해주기도 하지만 채점자의 신뢰도에 영향을 미칠 수 있다는 점을 제시하였다.

일반척도의 장점 같은 경우에는 여기서도 마찬가지로, 어... 스케일이

많은 경우에는 좀 세세하게 평가할 수 있다라는 게 장점이 되고. 반면에 또 이게 단점이 될 수도 있죠. 스케일이 많다는 것이. 전문적인 지식을 못 갖추고 있는 사람의 경우에는 방향이 좀 잘못 잡힐 수도 있으니까. (노, 인터뷰)

다수의 인터뷰 참여자들은 다른 평가 척도의 단점과 비교하여 세부과업별/개별 척도의 장점을 적극 제기하였다. 인터뷰 참여자들은 등급이 극히 협소한 이분척도와 지나치게 점수대가 많은 일반척도와의 중간지대로서의 개별척도의 장점을 부각시켰다.

채점하기에 정확한 채점표는..2번(개별척도)이요...이분척도는 Yes/No니까 이렇게 총점 매기기가 좀 어렵잖아요. 이거(2번)는 1, 2, 3으로 돼 있어서 일반척도의.. 비교적 1, 2, 3, 4, 5보다는 매기기도 그냥 쉽고. 이렇게 총점이 딱 이렇게 어느 정도 나오니까. (한, 인터뷰)

이거(일반척도)는 등급이 좀 너무 많아서 3점, 4점 사이에 어떤 걸 해야 할지 기준이 좀 불명확했는데, 이거(세부과업별 척도)는 편했던 것 같아요. 세 가지로 기준이 나와 있어서...(김, 인터뷰)

그밖에 세부과업별/개별 척도의 장점으로서, 해당과업에 대하여 자세하고 정확하게 측정할 수 있다는 점과 척도 기술이 문장단위로 되어 있어서 이해하기 쉽다는 점을 제시하였다. 그러나 ‘배’의 경우는 세부과업별 척도의 평가 항목이 세밀한 것을 측정하기 보다는 너무 포괄적인 사항을 측정하고 있어서 더 세분화된 평가항목이 필요하다는 입장을 제시하였다.

전 이게(세부과업별 척도) 그냥 가장 자세해서 좋았던 것 같아요. 지금 토픽마다 이렇게 해놓으신 거잖아요. 그래서 더 이렇게 자세하게 할 수 있어서... (황, 인터뷰)

편리했던 부분은.. 확실히, 2번 루브릭(세부과업별 척도)이 이게 확실히 문장으로 돼있으니까. 이런 부분만 맞다 안맞다 딱 보면 되는 거라서 편찮고. 이런 건 좀 저희가 확실히 채점하기가 편하고요. 또 일단 스케일이 작지 않습니까. 또 저희가 전문적으로 평가에 관해서 공부를 막 했던 사람들이 아니라서 스케일도 작고. 또 문장이 되게 구체적으로 제시가 돼 있으니까. 문장형식으로. (노, 인터뷰)

2번(세부과업별 척도)은 음... 요소가 좀... 너무... 크다고 그래야 되나요? 한번에 하나씩을 물으면 좀 더 정확하게 측정이 되겠는데. 그냥 큰 덩어리로 묻는 것 같아요. 예를 들어.. 이 항목은...더 있었으면 좋겠는데. (배, 인터뷰)

지금까지, 세 가지 평가 척도에 대한 학습자들의 인식을 조사하였다. 연구질문 2의 결과, 학생들은 개별척도의 장점에 대하여 더 적극적이며 민감하게 인식하고 있었으며 이분식 척도에 대해서도 긍정적인 의견을 피력하였다. 정리하자면, 세 가지 척도를 적절하게 혼용하여 쓴다면 평가의 타당성이 높아질 것이라는 ‘노’의 기대대로, 학습자들은 세 가지 척도가 각기 상호 보완적인 기능을 수행하고 있음을 인식하고 있음을 발견할 수 있었다.

4.3 전문인 채점자 측면에서 각 척도에 대한 인식은 어떠한가?

두 명의 전문인 평가자를 대상으로 사고표출기법과 인터뷰 자료를 수행하여 그 채점 과정을 분석한 결과, 전문인1은 상대적 비교에 의한 평가 방식을 수행한 반면, 전문인 B는 루브릭 기준에 충실히 대조하여 평가하였음을 알 수 있었다.

전문인1의 상대적 비교에 의한 평가 전략은 사고표출기법을 통해서 명확하게 관찰되었다. 그녀는 특정 수험자의 점수를 결정함에 있어서 이전의 수험자들의 수행능력과 비교하고 대조하였다. 그녀의 그런 채점 전략은 세 가지 유형의 척도에 관계없이 표출되었는데, 다른 척도보다 평가 항목이 많았던 이분식 척도에서 점수 결정에 어려움이 있었음을 토로하였다. 다만, 매우 높은 수준의 수험자의 에세이를 평가하는데 있어서는 이분법적 판단의 갈등이 적기 때문에, 이분식 척도의 활용이 매우 편리했음을 시사하기도 하였다.

수험자가 answer는 하고 있긴 한데.. 또 이런 부분이 그러니까 뭐. 뭐라 그래야 되지. 이게 내가 Yes라고 체크하면서도 그렇게 별로 하고 싶지 않은... 뭐 사실상 앞전에 채점 했던 학생들 중에서 상당히 내용도 좋고 괜찮았는데..(중간생략) 척도 C(세부과업/개별 척도)에서 1번 항목을 No라고 체크한 경우가 있는데. 그 부분은 참. No라고 체크하고 싶지 않았지만 No라고 체크했었던 경우가 있었는데. 그래서 척도 C가 약간... 좀 고민을 오히려 더 척도 A(일반적 척도)보다 많이 하게 만드는 경우가 생기네. (전문인1, 사고표출기법자료)

그리고 척도 B로 가서... 의견이 충실히 들어가 있고, 문제점도 있고. 해결책도... 해결책을 제시를 많이 했긴 한데, 문법적 실수나 비문이 눈에 많이 들어와서 3점을 주고 싶진 않지만 줘야 하나? 전에 부실하게 예시 없이 짧막하게 쓴 사람이 많아서 상대적으로 변별력을 주려고 한다면 3점을 줄 수 밖에 없을 거 같고. (전문인1, 사고표출기법자료)

이렇게 잘 쓴 에세이 같은 경우에는 척도C(이분식 척도)가 별로 고민될 게 없지. 왜냐면 Yes or No 이니까. 척도C같은 경우에는... 전부 Y를 줄꺼야. Yes. 척도 C도 만점이고. (전문인1, 사고표출기법자료)

사고표출기법을 통해서 드러난 전문인1의 평가방식은 그녀와의 인터뷰를 통해서도 재차 확인되었다. 그녀는 학생들과는 달리, 가장 평가하기에 효율적이고 편리한 척도로 개별척도를 지정하고, 제일 평가하기 어려웠던 척도로 이분식 척도를 내세우면서 협소한 등급으로 인한 불충분한 변별력을 상대적 평가를 통해서 해소하고자 했다는 의중을 밝혔다.

저는 이분식에서는 이게 너무 그게... 스케일이 작으니까. 거의 극단적이잖아요. 그래서 저는 약간 좀 상대적으로 주려고 했던 것 같아요. 그냥 그것만 안보구 내가 그거 실수의 정도에 차이가 있잖아요. 실수가... 얼마만큼 실수했는지 뭐 심했는지 적었는지 이런 게 있는데. 그래서 저는 좀... 고민을 진짜 많이 했어요. 이분식 할때는. 판애들 하고 비교하면서..(전문인1, 인터뷰)

위와 같은 전문가1의 상대적 비교에 의한 채점 방식이 일정부분 전문가1의 넓은 점수 폭(표 1의 표준편차 값 참조)과 일정부분 상관이 있는 것으로 보인다. 즉, 전문가1은 수험자간의 상호 비교에 의한 점수의 변별력을 하나의 중요한 가치로 삼았고, 그 결과로서 다른 채점자에 비하여 학생들 간 점수 차이가 가장 크게 발생된 측면이 있었다.

반면, 전문인2는 주로 루브릭에 제시되어 있는 평가 항목에 충실하여 평가 항목 제시 순서대로 수험자의 에세이를 검증하는 방식을 채택하였다. 전문인1과는 달리, 전문인2는 사고표출기법자료나 혹은 인터뷰를 통해서도 이전 학습자와 비교하여 점수를 결정했다는 흔적은 전혀 보이지 않았다. 하지만 전문인2 역시 전문인1과 척도별 특색에 대해서는 의견을 같이 하였는데, 가장 평가하기에 어려웠던 척도로 이분식 척도를 지적하고 정확한 채점이 가능한 척도로는 일반척도를, 가장 효율적이며 편리한 척도로는 세부과업별/개별척도라고 보고하였다. 다음 아래 자료는 전문인2의 평가과정을 잘 보여주고 있다.

국제결혼에 대한 자신의 의견이 충실히 들어가 있었고. 국제결혼에 대한 문제점, 해결책 또한 제대로 충실하게 설명을 하고 있고. 300자 내

외로 작성이 되었고 서론, 본론, 결론의 구조가 완벽하게 이루어져 있어요. 문단들여쓰기 및 구두법, 철자를 잘 지키고 있었고 문법 또한 잘 사용하였으며 상황에 맞는 적절한 어휘를 잘 사용하고 있네. 전반적으로 굉장히... 음... 사회전반적인 내용으로 접근을 하면서 이런 국제결혼에 대한 심도있는 내용을 다뤄준 내용이군. (전문인2, 사고표출기법자료)

연구자는 연구질문1의 평균과 표준편차 분석 결과, 전문인2가 두 가지 특이한 채점 양상들을 보이고 있음을 발견하였다. 그 특이한 채점 행태들 중 하나는 전문인2가 다른 채점자 그룹과 비교하여 점수의 변이를 가장 적게 주었던 채점자라는 것이었다. 따라서, 본 연구자는 점수 폭을 좁게 주었던 이유를 알기 위하여 전문인2에게 안전 지향적 채점 전략의 의식적인 사용 여부를 질문하였는데, 그녀는 그 사실을 강하게 부인하였다.

의식적으로 안전지향 적으로 채점을 해야 하겠다..그런 생각은 안 해봤어요. 전혀 그런 의도가 없었는데.. 전 주제를 보고 주제에 맞는 글은 잘 주려고 노력했고 그렇지 않은 글은 점수를 잘 주지 않으려고 했죠. (전문인2, 인터뷰)

전문인2의 두 번째로 흥미로운 채점 행태는 그녀가 모든 척도에 걸쳐서 가장 혹독한 평가경향을 보였다는 점이다. 그런 혹독한 평가 양식이 어떻게 발생되었는가에 대한 단서를 얻기 위하여 좀 더 심도 깊은 질문을 하였다. 그 질문에 대한 답변으로 그녀는 내용이나 구조 중심으로 비판적인 관점에서 평가를 수행하였다고 답변하였다.

저는 일관적으로 지금 문장간의 구조랑 내용을 좀 많이 봤거든요. 어휘량은... 학생들 대부분이 비슷비슷한 어휘를 사용을 했고. 숙어같은 경우에도 되게 비슷한 그런 구조를 갖고 있었어요. 근데 그런 서론이랑 본론이랑 결론... 그런 연결되는 부분을...잘 마무리를 지었는지. 그리고 거기에서 토픽이 그... 국제결혼에 대해서 본인이 생각하는 그런 문제점이라든가... 해결점을 제시를 해야 되는데. 그런 것을 예를 잘 못 들었더구나. 그리고 자기주장이 좀 불분명한 학생들은 좀 점수를 짜게 줬습니다. 비판적으로 채점을 해서 쳐 나간거죠. (전문인2, 인터뷰)

정리하자면, 전문인 채점자들은 척도별 특색에 대하여 학생들과는 다소 다른 견해를 표출하였다. 학생들이 가장 효율적이고 편리한 척도로서 이분식 척도를 제시한 반면, 전문인 채점자들은 세부

과업별/개별척도라고 응답하였다. 더 나아가 전문인 채점자들은 공통적으로 가장 채점하기 어렵고 힘들었던 평가척도로서 이분식 척도를 명시하였다.

5. 결론 및 시사점

본 논문의 연구 목적인 세 가지 척도별 비교를 위하여, 연구자들은 학생들과 두 명의 평가 전문가를 채점 과정에 참여시킨 후 그들의 채점자료와 설문자료, 인터뷰 자료를 수집하여 분석하였다. 혼합연구방법을 통해서 나온 연구 결과는 다음과 같이 정리할 수 있다.

정량적 분석을 통해 각 척도별 학생들과 채점 전문가들의 평가 점수를 비교한 결과, 채점자간 신뢰도는 세부과업별/개별척도를 활용한 평가체제에서 제일 높았고, 일반척도에서 그 신뢰도가 가장 낮았다. 특히 내용 영역 평가에서 자기-동료 평가 간, 전문인 평가 간의 채점자 신뢰도가 다른 척도에 비하여 현격히 높았음을 확인하였다. 성취율 비교를 통한 분석 결과는, 채점자들이 세부과업별/개별척도에 의한 평가에서 가장 후한 점수를 주는 경향을 보인 반면, 일반척도에서 혹독한 점수를 준 경향을 보였고, 평가영역에서는 내용이 다른 영역에 비하여 가장 낮은 점수를 기록하였다.

연구질문 1의 결과는 Knoch(2009)의 연구결과를 뒷받침하는 것으로서, 전문가의 선험적인 지식과 경험에 의하여 제작된 일반적 척도보다는 이전 수험자들의 에세이 분석과 의견 반영을 통하여 수정·제작된 세부과업별/개별척도에서 높은 신뢰도를 보였다고 할 수 있다. 신뢰도가 높기 위해서는 평가척도에 대한 채점자들의 명확한 이해가 담보되어야 한다는 점을 상기시키면(Hamp-Lyons, 1991), 실증적 방식으로 제작된 척도는 일선 교육현장에서 충분히 실행가능한 척도라고 보여진다.

설문지 분석과 인터뷰 분석을 통해서 수행된 학생들의 각 척도별 인식 조사 결과, 학생들은 채점하기 용이한 척도표로서 이분식 척도표를 지정하였으며, 영어작문 학습 활용에 적합한 척도와 정확한 점수 선정이 가능한 척도로서 세부과업별/개별척도를 압도적으로 선정하였다. 이 결과는 기존의 이분식 척도의 채점 편리성을 지적

한 이호(2010)와 황성삼(2008)의 연구결과와 일치한다고 할 수 있다.

연구질문 3과 관련하여 사고표출구술기법과 인터뷰 분석에 의한 전문인들의 인식 조사에서는, 두 명의 전문가들 모두 이분식 척도가 채점하기 가장 어렵다는 반응을 표출하였다. 반면에, 전문가들은 세부과업별/개별척도가 가장 평가하기 쉬운 척도로서 언급하였다. 실제로 연구질문 1에서의 신뢰도 분석에서 전문가들간 신뢰도가 이분식 척도에서 가장 낮게 나왔는데, 이는 전문인들이 40명의 수험자를 35개 항목에 걸쳐서 채점해야 하는 환경에서 기인한 측면도 있다. 마지막으로 전문인 채점자들은 수험자들의 상대비교에 의한 채점방식을 수행하거나 루브릭에 충실한 비판적 채점전략의 행태를 보였다. 하지만, 그런 전문인들의 평가행태는 세 가지 척도표에 상관없이 일관적으로 표출되었다.

위와 같은 연구 결과를 통해서, 몇 가지 중요한 시사점들이 도출될 수 있다. 우선 세부과업별/개별척도는 진단적 평가 상황이나 고부담시험의 수행 평가 상황에서 모두 적절하게 사용될 수 있는 평가척도라는 것이다. 특히 언어적 능력 이외에 내용적 지식을 측정하는 특수목적의 언어평가에서는 세부과업별/개별척도가 매우 현명한 선택이 될 수 있음을 본 연구 결과는 제시하고 있다. 다만, 이런 척도가 개별 과업, 개별 문항마다 제작되어야 한다는 어려움이 존재한다면, 언어나 구조영역의 평가 기준들은 고정을 시키되, 내용영역에 대한 평가기준들을 각 평가문항에 맞게 수정할 수 있는 방안을 생각해 볼 필요가 있다.

반면, 이분식 척도는 고부담시험보다는 진단평가 상황에서 더 적절하게 사용될 수 있는 평가척도이다. 한명의 전문인 채점자가 다수의 수험자를 정해진 시간 안에 평가해야 할 고부담시험의 평가환경에서, 한 수험자당 다수의 문항을 평가하는 것은 말 그대로 매우 큰 부담이 될 수 있고 이는 채점자내 신뢰도(intra-rater reliability)에 심각한 위협이 될 수 있다. 이런 이분식 척도의 단점은 영어작문 학습 상황에서는 장점으로 승화될 수 있는데, 다수의 평가항목을 가진 이분식 척도는 학생들이 자신의 실력에 대하여 스스로 그리고 체계적으로 검증을 받을 수 있는 토대를 제공한다는 점에서 진단 목적의 평가에 친화적인 척도라고 할 수 있다.

본 논문에서 제기된 심각한 문제점들 중의 하나는 일반적 척도에

의한 총괄평가가 총합평가에 비하여 채점자 신뢰도가 뚜렷하게 낮다는 점이다. 이 현상은 본 논문의 설계 오류에서 비롯된 것 일수도 있지만, 세부점수 영역을 합친 총합평가 방식이 총괄평가보다 신뢰도가 훨씬 높다면, 현재의 고부담 말하기·쓰기평가에서 전반적으로 시행되고 있는 총괄평가 방식에 대하여 재고할 필요가 있다. 총합평가 방식이 다소 이질적인 평가 세부구인들을 포괄하여 측정하는 교재평가나 기관평가와 같은 종합평가(evaluation)에서 최종결정 근거로 보편적으로 사용되고 있다는 점을 고려한다면, 수험자의 전반적인 능력에 대한 채점자의 주관성이 강하게 개입되는 총괄적 방식보다는 총합평가 방식이 평가체제의 정확성과 신뢰도를 높일 수 있는 방안이 될 수 있을 것이다.

본 논문은 실험의 내적 타당도와 관련하여 몇 가지 아쉬운 측면들이 존재한다. 우선 전문인 채점자간 신뢰도가 학생들의 자기-동료평가간 신뢰도에 비하여 전반적으로 낮은 점이 발견되었는데, 이는 전문인 채점자들이 자신의 채점 성향을 강하게 고집한 측면도 있지만 채점자 훈련이 연구자가 원하는 방향으로 제대로 구현되지 못했던 점도 존재한다. 학생들은 이미 평가자 훈련과 두 차례의 논술시험 평가를 통해서 본 논문의 쓰기 평가 과업에 대하여 익숙해졌고 또한 해당 척도에 대하여 충분하게 이해를 한 반면, 전문인들은 단 1회에 그친 평가자 훈련으로 인하여 해당평가 척도에 대한 충분한 이해가 부족할 수도 있었다는 점이다. 물론, 전문인들이 본 연구 이전에 말하기 쓰기 분야에 있어서 많은 채점 경험을 가지고 있었다 할지라도, 그들이 본 척도를 충분히 숙지했지 못했다면, 본 연구결과와 신뢰성에 위협이 될 수 있다. 따라서 추후 연구에서는 장기적 평가자 훈련 프로그램에 전문인 채점자들을 참여시켜 사용 척도에 대한 이해를 더 높이는 노력을 경주하도록 해야 할 것이다.

또한, 척도 간에 만점 기준이 다르고 척도 내에서도 세부 척도별로 등급이 다른 점으로 인하여 좀 더 심도 깊은 정량적 분석기법을 사용하지 못했던 측면도 있다. 예를 들어, GENOVA와 FACETS분석같은 측정학적 방식은 집단별로, 또한 개인별로 평가변수들에 대한 자세한 정보를 제공할 수 있지만, 세부평가요소들의 등급의 동등화가 이루어지지 못한 상태에서 이런 평가기법을 구현하는 것은 무리가 있었다. 하지만, 기술통계나 성취율같은 기본적인 통계자료

로도 채점자 정보에 대하여 부족하지만 유용한 정보를 제공할 수 있다고 본다.

지금까지 일반척도, 세부과업별/개별척도, 이분식 척도간의 비교를 통하여, 채점자로서의 학습자와 전문인 채점자가 각 척도에 대하여 어떤 채점 양상을 보이고 어떤 인식을 하고 있는지에 대하여 살펴보았다. Condon(2009)은 척도라는 것은 판단을 내리고 등급을 매기기 위한 단순한 수단이 아니라, 학습자의 학습을 촉진시키고 강화하는 촉매제로서의 역할도 존재함을 환기시켰다. 다양한 유형의 척도들을 교육현장 속에서 적절하게 사용하기 위해서는 평가 척도의 기능을 선발이나 학력측정에만 국한시키기 보다는, 수험자 측면에서 학습전략이나 영어 학습 동기와의 연관시키는 것으로까지 확대할 필요가 있다는 주장을 심각하게 고려하면서, 앞으로의 연구는 다양한 유형의 척도들에 대한 설계와 그 척도들의 타당도에 대한 충분한 고민을 담고 있어야 할 것이다.

참고문헌

- 이호. 2010. 귀납적 이분법 척도를 활용한 예비 영어교사의 영어말하기 평가 신뢰도와 효용성에 관한 연구. 《한국교원교육연구》 27, 215-237.
- 황성삼. 2008. EBB를 이용한 중등학교 말하기 평가척도 개발. 《영어교육》 63, 307-331.
- Carr, N. T. 2000. A comparison of the effects of analytic and holistic rating scale types in the context of composition tests. *Issues in Applied Linguistics* 11, 207-241.
- Chalhoub-Deville, M. 1995. Deriving oral assessment scales across different tests and rater groups. *Language Testing* 12, 16-33.
- Charney, D. 1984. The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English* 18, 65-81.
- Condon, W. 2009. Looking beyond judging and ranking: Writing assessment as a generative practice. *Assessing Writing* 14, 141-156.
- Dörnyei, Z. 2007. *Research methods in applied linguistics*. London: Oxford University.
- Douglas, D. 2000. *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.
- East, M. 2009. Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing* 14, 88-115.
- Fulcher, G. 1996. Does thick description lead to smart tests? A

- data-based approach to rating scale construction. *Language Testing* 13, 208-238.
- Green, A. 1998. Verbal protocol analysis in language testing research: *A handbook*. Cambridge: Cambridge University Press.
- Hamp-Lyons, L. 1991. Scoring procedures for ESL contexts. In L. Hamp-Lyons, ed., *Assessing second language writing in academic contexts*, 241-278. Norwood, NJ: Ablex.
- Hwang, S. S. 2010. *Discourse analysis of EFL interview test at secondary school: A focus on oral fluency*. Unpublished doctoral dissertation. Chung-Ang University, Korea.
- Kim, Y. H. 2010. *An argument-based validity inquiry into the Empirically-Derived Descriptor-Based Diagnostic (EDD) assessment in ESL academic writing*. Unpublished doctoral dissertation. University of Toronto, Canada.
- Knoch, U. 2009. Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing* 26, 275-304.
- Lane, J. L. 2008. *The basics of rubrics*. Retrieved July 20, 2011, from the World Wide Web: <http://www.schreyerinstitution.psu.edu/pdf/RubricBasics.pdf>
- McGriff, S. J. 2007. *The effect of an analytic trait, task-specific rubric with or without explanation on self-efficacy and academic achievement*. Unpublished doctoral dissertation. The Pennsylvania State University.
- Moskal, B. M. 2000. Scoring rubrics: what, when and how?. *Practical Assessment, Research & Evaluation* 7. Retrieved July 21, 2011 from <http://PAREonline.net/getvn.asp?v=7&n=3>
- Song, D. Y. 2009. Applicability of an EBB scale to English speaking performance assessment in the Korean secondary school. *English Language Teaching* 21, 111-135.
- Upshur, J. and C. Turner. 1995. Constructing rating scales for second language tests. *English Language Teaching Journal* 49, 3-12.
- Upshur, J. and C. Turner. 1999. Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing* 16, 82-111.
- Weigle, S. C. 2002. *Assessing writing*. Cambridge, UK: Cambridge University Press.
- Zimmaro, D. M. 2004. *Developing grading rubrics*. Retrieved July 20, 2005 from the University of Texas at Austin, Division of Instructional Innovation and Assessment Web site: <http://www.utexas.edu/academic/mec/research/pdf/rubricshandout.pdf>

이 호

서울시 동작구 흑석동 221번지

중앙대학교 영어교육학과

Tel: (02) 820-5393 / 010-2354-4479

E-mail: holee@cau.ac.kr

Homepage: <http://www.tesol.pe.kr>

김혜영

서울시 동작구 흑석동 221번지

중앙대학교 영어교육학과

Tel: (02) 820-5399

E-mail: englishnet@cau.ac.kr

이진화

서울시 동작구 흑석동 221번지

중앙대학교 영어교육학과

Tel: (02) 820-5400

E-mail: jinhlee@cau.ac.kr

논문접수: 2011.07.31

논문수정: 2011.09.05

게재결정: 2011.09.15

부록 1

일반척도

채점제안: 세부채점영역의 두 군데에서만 (혹은 그 이상) 1 점에 해당되면, 그 해당 채점영역은 1 점.

채점영역	세부채점영역	해당 사항
총괄영역 holistic	구 조	아이디어가 전혀 조직화되어 있지 않고 불충분한 내용이다. 전혀없는 접속사.
	내 용	주제문에 대한 보충설명이 전혀 혹은 거의 없거나 상관없는 이야기만 나열.
	어휘 사용	문법의 오용이 두드러져서 의미를 알수 없을 정도이다. 심각한 구두법의 문제. 문장의 형태가 극히 단조롭고 전체적으로 무엇을 이야기 하는지 알 수 없다.
구 조 Organizational	서론, 본론, 결론	서론, 본론, 결론 구조가 없거나 서론, 본론, 결론중에 하나 이상이 빠져있다.
	문장간의 연결성	문장과 문장을 연결하는 접속사가 거의 없다. 문단내 글의 흐름이 끊겨 메시지를 알 수 없다. 주제문(thesis statement)가 전혀 없다.
	문단간의 연결성	의미가 서로 연결되지 못할 정도고, 문단과 문단을 연결하는 접속사가 없다.
내 용 content	제시된 토픽준수	제시된 토픽과는 거리가 멀다.
	구체적예증제시	구체적인 증거나 예시가 전혀 제시되어 있지 않다.
	자신의 주장표현	자신의 주장이 전혀 보이지 않고, 읽은 article 을 요약하기만 했다.
	무의미한반복	문장안에서 불필요한 단어 사용이나, 문단내 특정 단어의 남용이 심각하다. 문장형태도 똑같은, 같은 의미를 가진 문장들이 자주 출몰한다.
	배껴쓰기	대부분 Article 에 적합한 문장이나 의미를 원작자 표시없이 그대로 배껴썼다.
어휘사용 language use	문법의 정확성	거의 모든 분야에서 문법상 오류를 저지르고, 의미전달 막는 오류 투성이다.
	단어의 철자법	15 개 이상의 단어들의 철자법이 올바르게 쓰여지지 못했다.
	적절한 단어사용	문장의 흐름상 전혀 의미가 맞지 않는 단어, 속어가 각 문단마다 네다섯개 이상 쓰였다. 따라서 글의 의미를 이해하기가 힘들다.
	다양한문장구조	제대로 끝나지 못한 문장들이 대다수이고 간단한 단문만 나타나고 있다.

부록 2

세부과업별/개별 척도

채점기준	채점요소	점수
내용	1. 국제결혼에 대한 자신의 의견이 충실히 들어가 있다.	1 2 3
	2. 국제결혼이 초래할 문제점(단점 및 우려할 점)이 충실히 들어가 있다.	1 2 3
	3. 그 문제점에 대한 해결책이 조리있게 제시되어 있다.	1 2 3
	300자 내외로 작성되었다.	1 2
구조와 어휘 사용	서론결의 구조가 제대로 이루어져 있다.	1 2 3
	문단들여쓰기 및 구두점, 철자를 잘 지키고 있다	1 2
	올바른 문법을 사용하고 있다.	1 2 3
	상황에 맞는 적절한 어휘를 사용하고 있다.	1 2 3
	같은 문장 구조의 반복적 나열이 적고, 다양한 문장구조를 사용한다.	1 2 3

부록 3

이분식 평가 척도 [Kim(2010), p. 267에서 인용]

평가항목	점 수
1. This essay answers the question.	☐ Y ☐ N
2. This essay is written clearly enough to be read without having to guess what the writer is trying to say.	☐ Y ☐ N
3. This essay is concisely written and contains few redundant ideas or linguistic expression	☐ Y ☐ N

4. This essay contains a clear thesis statement.	☐ Y ☐ N
5. The main arguments of this essay are strong.	☐ Y ☐ N
6. There are enough supporting ideas and examples in this essay.	☐ Y ☐ N
7. The supporting ideas and examples in this essay are appropriate and logical.	☐ Y ☐ N
8. The supporting ideas and examples in this essay are specific and detailed.	☐ Y ☐ N
9. The ideas are organized into paragraphs and include an introduction, a body, and conclusion.	☐ Y ☐ N
10. Each body paragraph has a clear topic sentence tied to supporting sentences.	☐ Y ☐ N
11. Each paragraph presents one distinct and unified idea.	☐ Y ☐ N
12. Each paragraph is connected to the rest of the essay.	☐ Y ☐ N
13. Ideas are developed or expanded well throughout each paragraph.	☐ Y ☐ N
14. Transition devices are used effectively.	☐ Y ☐ N
15. This essay demonstrates syntactic variety, including simple, compound, and complex sentence structures.	☐ Y ☐ N
16. This essay demonstrates an understanding of English word order.	☐ Y ☐ N
17. This essay contains few sentence fragments.	☐ Y ☐ N
18. This essay contains few run-on sentences or comma splices.	☐ Y ☐ N
19. Grammatical or linguistic errors do not impede comprehension.	☐ Y ☐ N
20. Verb tenses are used appropriately.	☐ Y ☐ N
21. There is consistent subject-verb agreement.	☐ Y ☐ N
22. Singular and plural nouns are used appropriately.	☐ Y ☐ N
23. Prepositions are used appropriately.	☐ Y ☐ N
24. Articles are used appropriately.	☐ Y ☐ N
25. Pronouns agree with antecedents.	☐ Y ☐ N
26. Sophisticated or advanced vocabulary is used.	☐ Y ☐ N
27. A wide range of vocabulary is used.	☐ Y ☐ N
28. Vocabulary choices are appropriate for conveying the intended meaning.	☐ Y ☐ N
29. This essay demonstrates facility with appropriate collocations.	☐ Y ☐ N
30. Word forms (noun, verb, adjective, adverb, etc) are used appropriately.	☐ Y ☐ N
31. Words are spelled correctly.	☐ Y ☐ N
32. Punctuation marks are used appropriately.	☐ Y ☐ N
33. Capital letters are used appropriately.	☐ Y ☐ N
34. This essay contains appropriate indentation.	☐ Y ☐ N
35. Appropriate tone and register are used throughout the essay.	☐ Y ☐ N